MATHEMATICAL PROGRAMMING METHODS OF PATTERN CLASSIFICATION

By

Richard C. Grinold

June 1969

DDC

OCT 14 1969

C

Technical Report No. 591

26

MATHEMATICAL PROGRAMMING METHODS

OF PATTERN CLASSIFICATION

By

Richard C. Grinold

Technical Report No. 591

June 1969

Division of Engineering and Applied Physics

Harvard University · Cambridge, Massachusetts

MATHEMATICAL PROGRAMMING METHODS
OF PATTERN CLASSIFICATION

By

Richard C. Grinold

Division of Engineering and Applied Physics

Harvard University   Cambridge, Massachusetts

ABSTRACT

This paper studies four mathematical programming methods
which are useful in pattern classification. Two of the models are for
linearly separable problems, while the others work without separability.

INTRODUCTION

This report is designed to supplement "On Pattern Classification-Introduction and Survey,"[*][7], by describing several mathematical programming approaches to the classification problem. We'll assume that the reader is familiar with the Ho and Agrawala paper (at least sections I, II, and IV) and draw on the motivation, notation, and definitions used there.

Four mathematical programming models are described in detail, and two more are mentioned briefly. Others exist, and are referenced in the publications cited here. The four models were selected for their computational and conceptual properties.

Before describing the contents of the paper, we'll expand upon and change some of the notation adopted in [7].

Definitions:

(i). Instead of $x^1(i)^T$ and $x^0(j)^T$, the training samples from classes one and zero will be denoted by m component row vectors A_i^1 and A_j^0.

(ii). For $k = 0, 1$; A^k is the $n_k \times m$ matrix whose rows are A_i^k, $i = 1, 2, \ldots, n_k$

(iii). h, ℓ, e, q, and f are vectors of ones. There dimensions are given below.

$h \sim n_1 \times 1$; $\ell \sim n_0 \times 1$; $e \sim m \times 1$; $q \sim n_1 \cdot n_0 \times 1$ and $f \sim n \times 1$, where $n = n_1 + n_0$.

(iv).   A <u>linear decision function</u> is an $(m+1) \times 1$ vector $w = \begin{bmatrix} \lambda \\ u \end{bmatrix}$.

Using these definitions we note that

$$A = \begin{bmatrix} h & A^1 \\ -\ell & -A^0 \end{bmatrix}$$

Suppose the patterns are described by an m component row vector x, then the decision function defined by w is

$$f(x) = \lambda + xu = (1, x)w.$$

(v).   A linear decision function w is a <u>separator</u> if

$$Aw > 0$$

Problems are specified by their range of attributes.
The ranges are defined below.

(vi). $S^1 = \{x | P(x | H^1) > 0\}$
        $S^0 = \{x | P(x | H^0) > 0\}$

(vii).   The operator C will denote convex closure. Thus $C(S^1)$
        is the closed convex hull of S^1.

(viii). The problem is <u>separable</u> if $C(S^0)$ and $C(S^1)$ are disjoint.
        If they intersect the problem is <u>nonseparable</u>.

(ix).   The problem is <u>decidable</u> if S^0 and S^1 are disjoint.

(x).   We shall define $C(A^k)$ as the convex hull of the rows of
        A^k, $k = 0, 1$.

Thus

$$C(A^0) = \{b | b = zA^0, \; z\ell = 1, \; z \geqq 0\}$$

$$C(A^1) = \{b | b = yA^1, \; yh = 1, \; y \geqq 0\}$$

Five sections and an appendix follow. Sections one and two describe models used in the separable and nonseparable cases. The third section remarks on the model's flexibility in terms of accommodating new data and use in judging new features. Section four considers the model's generalization properties, while the last section describes an application. The appendix is a brief introduction to linear and quadratic programming.

Of the four models, two have been described in the published literature. One of the unpublished models is due to Canon and Cullum [3], the other is the author's responsibility [6].

I. THE SEPARABLE CASE

Charnes [4] and Mangasarian [9] independently proposed a linear programming model for separating disjoint polyhedrons. Their distinct approaches illustrate the duality principle of linear programming. Charnes asks if the sets $C(A^0)$ and $C(A^1)$ are disjoint, while Mangasarian looks directly for a separating hyperplane.

By definition, the $C(A^k)$ will intersect if and only if the system (1) has a feasible solution.

$$zA^1 - yA^0 = 0, \quad yh = 1, \quad z\ell = 1, \quad y \gtreqless 0, \quad z \gtreqless 0 \tag{1}$$

We can discover a solution of (1) by adding artifical variables to the system and minimizing the infeasibility. This gives us a linear program.

$$\text{Minimize } (r + s)e \tag{2}$$

Subj. to

$$zA^0 - yA^1 + rI - sI = 0$$

$$z\ell \qquad\qquad = 1$$

$$\qquad yh \qquad = 1$$

$$z \gtreqless 0, \quad y \gtreqless 0, \quad r \gtreqless 0, \quad s \gtreqless 0$$

This problem has $m + 2$ equality constraints with $n + 2m$ nonnegative variables. The value, $(r + s)e$, is nonnegative since r and s are nonnegative. Finally, we can easily construct a first basic feasible solution of (2).

The alternate approach involves the decision function directly. Suppose the m vector u and scalars (γ, δ) satisfy the following conditions:

$$\delta - \gamma > 0 \tag{3}$$
$$-A^1 u - h\gamma \leqq 0$$
$$A^0 u + \ell\delta \leqq 0$$

then $w = (\frac{\gamma + \delta}{2}, u)$ is a separator. A solution of (3) can be discovered by solving (4).

$$\text{Maximize } \delta - \gamma \tag{4}$$

Subj. to

$$-A^1 u - h\gamma \leqq 0$$
$$+A^0 u + \ell\delta \leqq 0$$
$$-e \leqq u \leqq e$$

Problem (4) has n inequality constraints, two free variables (γ, δ), and m variables with upper and lower bounds. The bounds-rule out infinite solutions. Evidently, $(\gamma, \delta, u) = (0, 0, 0)$ is a feasible solution of (4).

Appealing to the results in the appendix we can state that problem (4) is the dual of problem (2), and the duality theorem applies. This guarantees the existence of optimal solutions $(\bar{z}, \bar{y}, \bar{r}, \bar{s})$ and $(\bar{\gamma}, \bar{\delta}, \bar{u})$ such that:

$$(\bar{r} + \bar{s})e = \bar{\delta} - \bar{\gamma} \geqq 0.$$

There are two possibilities. If $\delta - \gamma > 0$, then $(\frac{\bar{\gamma} + \bar{\delta}}{2}, \bar{u})$ is a separator. If $e(\bar{r} + \bar{s}) = 0$, then $(\bar{z}, \bar{y})$ solves (1), and the convex hulls intersect. These facts are summarized below.

Theorem: (5)

    (i).   Problems (2) and (4) have optimal solutions with equal, nonnegative values.

(ii). If the optimal value is zero, the patterns are not linearly separable.

(iii). If the optimal value is positive, then $\bar{w} = (\frac{\bar{\gamma} + \bar{\delta}}{2}, \bar{u})$ defines a separator which maximizes

$$\text{Min}[A_i w \mid i = 1, 2, \ldots, n]$$

Subj. to

$$-1 \leq w_j \leq 1, \quad \text{for} \quad j = 1, 2, \ldots, m$$

Statements (i), (ii), and the first part of (iii) are established above. The final statement can be established by contradiction.

The linear programs will be solved using some variant of Dantzig's, [2], simplex method. This is a rapidly convergent combinatorial procedure, while the adaption algorithms, see [7] Table I, are gradient descent techniques which converge slowly. If the patterns are not separable, slow and nonconvergence can be confused. See [9], pg. 451, for a more detailed comment along this line. The adaption algorithms do have the advantage of simplicity, but this is largely offset by the wide availability of professionally written linear programming codes. Either (2) or (4) can be solved, but the simplex algorithm is more efficient with fewer nontrivial constraints. It is not very sensitive to the number of variables. Since $m + 2 \ll n$ it is reasonable to solve (2).

Canon and Cullum [3] have proposed a quadratic programming method for the separable case. Although it is generally more difficult to solve quadratic programs, the authors take advantage of the problem's special structure and claim their method is competitive with the linear programming model.

For each i and j, $i = 1, 2, \ldots, n_1$, $j = 1, 2, \ldots, n_0$; we can define a difference vector

$$D_k = A_i^1 - A_j^0 \quad \text{for} \quad k = 1, 2, \ldots, n_1 n_0.$$

The vectors D_k are the rows of the $n_1 n_0 \times m$ matrix D. Recall $C(D) = \{u \mid u = yD, \; yg = 1, \; y \geq 0\}$. It is easy to establish that $C(A^0)$ and $C(A^1)$ will be separable if and only if the origin is not contained in $C(D)$.

This suggests a test for separability: find the vector in $C(D)$ with minimum norm. The problem can be written in two ways:

$$\text{Minimize} \quad \frac{uIu'}{2} \tag{6}$$

Subj. to

$$yD - uI = 0$$
$$yg \quad\quad = 1$$
$$y \geq 0$$

$$\text{Minimize} \quad \frac{yDD'y'}{2} \tag{7}$$

Subj. to

$$yg \quad\quad = 1$$
$$y \geq 0$$

The following facts about (6) and (7) should be clear: they are equivalent, the objectives are convex and quadratic, they have optimal solutions with nonnegative values, and the sets are separable if and only if the optimal value is positive.

It is well known that any point in $C(D)$ can be expressed as a convex combination of at most $m + 1$ rows of D. This fact is used to reduce the problem's size. The algorithm solves a modified version of (6), restricting

attention to a subset of m + 1 rows. A test sees if the restricted solution is optimal with all rows considered. If so, (6) is solved. If not, a new row is added, an old row dropped, and the algorithm proceeds, finding an optimal solution in a finite number of steps. The optimal solution of (6) defines the linear decision surface.

Suppose $(\bar{y}, \bar{u})$ solves (6), $\bar{u} \neq 0$, and

$$\gamma = \text{Min } \{A_i^{1}\bar{u}' \mid i = 1, 2, \ldots, n_1\}$$

$$\delta = \text{Max } \{A_j^{0}\bar{u}' \mid j = 1, 2, \ldots, n_0\}$$

then $(\frac{\gamma + \delta}{2}, \bar{u})$ is a separator. If $\bar{u} = 0$, no separator exists. This is demonstrated in the appendix using the Kuhn-Tucker theorem. Canon and Cullum do the same by showing problem (6) is equivalent to:

$$\text{Max } [\text{Min } \{uz \mid u\epsilon C(D)\}] \tag{8}$$

Subj. to

$$z'Iz \leq 1$$

II. NONSEPARABLE

One approach to the nonseparable case was taken in [6]. A description will require two definitions.

(xi). Let $a = \sum_{i=1}^{n} \frac{A_i}{n}$ be the average of the rows of A.

(xii). For any decision function w, let the quality of w be defined as

$$\text{Min } [A_i w \mid i = 1, 2, \ldots, n]$$

If w is a separator, the quality is positive. If w is not a separator, the negative of the quality (a nonnegative number) measures the largest error the decision surface makes. To obtain a decision surface of highest quality we solve

$$\text{Maximize } \{ \text{Min}[A_i w \mid i = 1, 2, \ldots, n]\}$$

Subj. to $\quad\quad\quad aw = 1$

The constraint is a normalization.

This problem can be transformed into a linear program by introducing a new variable ρ and requiring $\rho \leqq A_i w$ for $i = 1, 2, \ldots, n$. The new problem and its dual are given below.

Maximize ρ $\hspace{6cm}$ (9)

Subj. to

$$Aw - f\rho \geqq 0$$
$$aw \quad\quad = 1$$

Minimize γ $\hspace{6cm}$ (10)

Subj. to

$$yA - \gamma a = 0$$
$$yf \quad\quad = 1$$
$$y \geqq 0$$

Problem (10) has $m + 2$ equality constraints, n nonegative variables, and one free variable.

The main result of [6] is:

Theorem $\hspace{8cm}$ (11)

(i). Problems (9) and (10) have optimal solutions with equal objective values iff and a $\neq$ 0. When a = 0, (9) is infeasible.

(ii). If (w, ρ) solves (9) and $\rho > 0$, then w defines a separator of maximum quality.

(iii). If (w, ρ) solves (9) and $\rho \leqq 0$, then the patterns are not separable and w defines a decision surface that minimizes the maximum error.

This is equivalent to (5) in the separable case. In addition, a meaningful decision surface is generated if the patterns are not separable.

Smith, [13], has another approach. Note that $Aw > 0$ has a solution iff $Aw \geqq f$ has a solution. In this spirit, we can solve

$$\text{Minimize } f'v$$

Subj. to
$$Aw + Iv \geqq f$$
$$v \geqq 0$$

The v_i's measure the size of any error in the classification of the ith sample. Thus if $A_i w \geqq 1$, there is no error and $v_i = 0$. If $A_i w < 1$, v_i is positive. There is some difficulty if $0 < A_i w < 1$. In this case the pattern is correctly classified, but an error is counted. This behavior is observed in optimal solutions.

The dual, (12), is a linear program with $m + 1$ equality constraints and n nonnegative variables with upper bounds. It is relatively easy to solve, [5].

$$\text{Maximize } yf \tag{12}$$

Subj. to
$$yA = 0$$
$$0 \leqq y \leqq f'$$

This model suggests several conceptually interesting but computationally difficult variations. For instance, we could minimize the sum of squared errors. This leads to a quadratic program:

$$\text{Minimize } v'Iv$$

Subj. to

$$Aw + Iv \geqq f$$

$$v \geqq 0$$

Another variant maximizes the number of correctly classified samples:

$$\text{Maximize } \sum_{i=1}^{n} \delta(A_i w)$$

Subj. to $-1 \leqq w_i \leqq 1 \quad i = 0, 1, 2, \ldots, m$

$\delta(\cdot)$ is the step function; one if its argument is positive, zero otherwise. This problem can be reformulated as an integer program, [12] pp. 194-8.

Another method of treating the nonseparable case was proposed by Mangasarian, [10]. The approach is similar to Arkadev and Braveman [1], i. e. a piecewise linear decision surface is created which decides correctly about all the data. Mangasarian uses mathematical programming to construct the decision surface. We will not examine that algorithm in detail, but we do comment on its generalization properties in section 4.

III. FLEXIBILITY

This section examines the ability of the different models, (2), (6), and (12) to handle new data and yield information useful in selecting new

features. Models (2), (10), and (12) can accept new data points and find a new decision surface easily. In each case, adding a new point is equivalent to introducing a new activity (column) into the linear program.

Model (6) has a similar property. For example, suppose a new point in class one, $A^1_{n_1+1}$ is observed. This adds n_0 new rows to the matrix D. If

$$A^1_{n_1+1} u' \geq \text{Min} [A^1_{n_i} u' \mid i = 1, 2, \ldots, n_1]$$

no change is needed, the old decision surface is still optimal. If the inequality does not hold, we continue to apply the Canon-Cullum alg$_{\;}$ithm until a new optimal solution is obtained.

Introducing a new feature in (2), (10), or (12), [†] adds a new constraint (row) to the linear program. If several new features are being considered, we can devise a heuristic rule for chosing among them. Try the current optimal solution for each new constraint. Select the constraint which is the furtherest from being satisfied. If the optimal solution satisfies all the new constraints, it is still optimal. This selects the feature which maximizes the rates of improvement of the solution. Then a new optimal solution can be obtained using the dual simplex method.

[†] There doesn't seem to be any way that new feature can be accommodated by model (6).

IV. GENERALIZATION

The generalization properties of the models are examined in this section. In particular, we are interested in the decision surfaces generated as the number of sample points n becomes large. For each n the models produce a decision surface defined by a nonzero m + 1 vector. Without loss of generality we can uniformly bound these vectors. Thus, there will be subsequences which converge. We shall study the properties of the limiting decision surface.

For example, assume $C(S^1)$ and $C(S^0)$ are disjoint with one set compact, and consider model (2). Let (λ^n, u^n) be the normalized optimal decision surface for the n sample problem and let (λ, u) be a limiting surface: i.e. $(\lambda^n, u^n) \to (\lambda, u)$ on some subsequence. The following theorem asserts (λ, u) is optimal for the limiting problem.

Theorem:

With probability one (wp· 1) there exists a $\rho > 0$ such that (λ, u, ρ) solve:

$$\text{Maximize } \rho$$

Subj. to

$$\lambda + xu - \rho \geqq 0 \qquad x \in C(S^1)$$

$$-\lambda - xu - \rho \geqq 0 \qquad x \in C(S^0)$$

$$-e \leqq u \leqq e$$

Proof:

There exists a hyperplane which strictly separates $C(S^1)$ and $C(S^0)$. Therefore the problem has an optimal solution with positive value.

Suppose (λ, u) and some $\rho > 0$ are not optimal. A contradiction can be established by appealing to the the facts that (λ, u) is (2) feasible for all n, and that (λ, u) is the limit of a subsequence of optimal solutions.

Three comments are in order. First it is obvious that similar results hold for models (6), (10), and (12). Secondly, if compactness is dropped a weaker, $\rho \geq 0$, statement is true. Finally, if separability doesn't hold, then (wp, 1) all models will indicate this for some large value of n.

Assuming decidability we could obtain a like result using the piecewise approach, [10]. Additional regularity assumptions are needed to allow a piecewise linear function defined by a finite number of hyperplanes. Without decidability, the piecewise approach would struggle in vain to produce a perfect decision function.

Model (10) will work in the separable case, but it has questionable generalization properties. It is very sensitive to the tails of the distributions. The decision surface minimizes the maximum error, therefore it will react to the worst points or prehaps to a faulty observation. Things can get worse.

Let a^k be the finite means of the distributions, $P(x | H^k)$ for k = 0, 1. Then the row average of A will converge (wp. 1) to

$$a = \begin{pmatrix} P(H^1)\, a^1 - P(H^0)\, a^0 \\ P(H^1) - P(H^0) \end{pmatrix}$$

The following is an example of what can go wrong. Suppose $P(H^1) > P(H^0)$, and the sets described below have a nonvoid intersection:

$$L = \{d \mid d = \gamma a, \ \gamma \leqq 1\}$$

$$Z = \{d \mid d = \begin{pmatrix} b \\ -1 \end{pmatrix}, \ -b \ \epsilon \ C(S^0)\}$$

then the limiting optimal solution is given by $w = (\dfrac{1}{P(H^1) - P(H^0)}, \ 0)$ i. e. the decision function is

$$f(x) = \frac{1}{P(H^1) - P(H^0)} > 0 \quad \text{for all } x$$

The fact that f is correct more than not offers little consolation. Note that this phenomenon will occur if $S^0 = S^1 = R^m$, and $P(H^0) \neq P(H^1)$: e. g. multivariate normal.

The generalization properties of (12) seem to be the best. It is a reasonable conjecture that the limiting decision surfaces of (12) are optimal solutions to the following:

Minimize F(w)

Subj. to

$$-1 \leqq w_i \leqq 1 \quad i = 0, 1, 2, \ldots, m$$

where

$$F(w) = P(H^1) \int_{X^0} (-\lambda - ux) P(x \mid H^1) \, dx + P(H^0) \int_{X^1} (\lambda + ux) P(x \mid H^0) \, dx$$

is the expected error distance. It is also reasonable to assume that the limiting decision surfaces of the integer program mentioned in section two will minimize the probability of error among all linear decision functions. A brief attempt was made to prove these conjectures, but the proof is elusive.

V. EXAMPLE

Models (10) and (12) were employed to design decision functions using data from a NASA biomedical experiment. Two types of electro-encephalograms (brainwaves, EEG) were recorded. In one instance the subject was watching a strobe light. In the other case the light was not visible. The object is to distinguish the two cases using the EEG data.

Of a possible one hundred features K. Prahbu selected five, using a distance-dispersion technique and prepared the data for the linear programming models. The parameters were $n_0 = 165$, $n_1 = 155$, $m = 5$, $n = 320$, and the problems were solved on an IBM 360-65 using the mathematical programming package, MPS 360, [11]. Results are tabulated below.

Model (10)   Solution Time 0. 09 min.

Errors	Number	Percentage
Type I	25	16. 6
Type II	21	12. 7
Total	46	14. 4

Model (12)   Solution Time 0. 92 min.

Errors	Number	Percentage
Type I	18	11.6
Type II	24	14. 5
Total	42	13. 1

Notice the performance of model (12) is slightly better although the solution time is longer. Both problems had unique optimal solutions and 31 of the points were incorrectly classified by both techniques.

APPENDIX

Linear and Quadratic Programming

Several results from mathematical programming have been used in this report. This appendix attempts to motivate and explain these results while citing more substantial references.

A <u>linear program</u> is an optimization problem

$$\text{Min} \sum_{j=1}^{m} x_j c_j$$

Subj. to

$$\sum_{j=1}^{m} x_j a_{ji} = b_i \qquad i = 1, 2, \ldots, n$$

$$x_j \geqq 0 \qquad j = 1, 2, \ldots, m$$

Our vector notation is

$$\text{Min } xc$$

Subj. to

$$xA = b$$

$$x \qquad 0$$

c is $m \times 1$, A $m \times n$, and b $1 \times n$. We shall call this problem the <u>primal</u>. There is an associated <u>dual</u> problem:

$$\text{Max} \quad by$$

Subj. to

$$Ay \leqq c$$

Linear programs appear in many forms: maximization or minimization, equalities or inequalities, nonnegative or unrestricted variables. Any problem can be transformed into the same form as our primal, which allows us to know its dual. The dual can be found directly using the diagrams on pp. 126-7 of [2].

An efficient algorithm known as the simplex method, has been devised to solve linear programs. In a finite number of steps it finds a feasible solution (if one exists), then again in a finite number of steps it determines an optimal or an unbounded solution. An optimal dual solution is supplied as a by product of the calculations.

The principle theoretical result in linear programming relates primal and dual.

Theorem: [2] pg. 129

If both primal and dual have feasible solutions, they have optimal solutions (x, y) such that

$$xc = by$$

We shall discuss quadratic programming in the context of problem (6).

$$\text{Min } \frac{uIu'}{2}$$

Subj. to

$$yD - uI = 0$$
$$yg \quad = 1$$
$$y \quad \quad 0$$

A central result in the study of these problems is the Kuhn-Tucker theorem, [8]. In our case it states:

Theorem:

(y, u) is optimal for (6) if and only if there exist (x, z, λ) such that

$$g\lambda + Dx + z = 0$$
$$u' - x \qquad = 0$$
$$y \geqq 0$$
$$z \geqq 0$$
$$yg = 1$$
$$yD - uI = 0$$
$$yz = 0$$

Suppose $u \neq 0$ is optimal in (6), then $uIu' > 0$. Juggling the above equations we can easily establish that

$$\lambda = -uIu' < 0$$

and

$$Du' \geqq g(uIu') > 0.$$

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Division of Engineering and Applied Physics Harvard University Cambridge, Massachusetts	Unclassified
	2b. GROUP

3. REPORT TITLE

MATHEMATICAL PROGRAMMING METHODS OF PATTERN CLASSIFICATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Interim technical report

5. AUTHOR(S) (First name, middle initial, last name)

Richard C. Grinold

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
June 1969	25	13

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0298-0006 & NASA Grant	
b. PROJECT NO. NGR 22-007-143	Technical Report No. 591
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; it distribution is unlimited. Reproduction in whole or in part is permitted by the U.S. Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research

13. ABSTRACT

This paper studies four mathematical programming methods which are useful in pattern classification. Two of the models are for linearly separable problems, while the others work without separability.

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Linear Programming						
Quadratic Programming						
Pattern Classification						
Pattern Recognition						
Non Parametric Hypothesis Testing						

REFERENCES

1. Arkadev, A. G. and Braveman, E. M., Computers and Pattern Recognition, Washington, D. C., Thompson 1967.

2. Dantzig, G. B., Linear Programming and Extensions, Princeton University Press, Princeton, N. J., 1963.

3. Canon, M. D. and Cullum, C. D., "The Determination of Optimum Separating Hyperplanes, I: A Finite Step Procedure," Report (un-numbered) of IBM Watson Research Center, Yorktown Heights, New York.

4. Charnes, A., "Some Fundamental Theorems of Perception Theory and Their Geometry," in Computer and Information Sciences, J. T. Tou and R. H. Wilcox (eds.) pp. 67-74, Spartan Books, Washington, D. C. 1964.

5. Grinold, R. C., "A Comment on 'Pattern Classifier Design by Linear Programming,'" IEEE Transactions on Computers, to appear.

6. Grinold, R. C., "A Note on Pattern Separation," submitted to Operations Research.

7. Ho, Y. C. and A. K. Agrawala, "On Pattern Classification Algorithms - Introduction and Survey," Proc. IEEE Vol. 56, No. 12, December 1968.

8. Kuhn, H. W. and A. W. Tucker, "Nonlinear Programming," Econometrica, Vol. 19, No. 1, January 1951.

9. Mangasarian, O. L., "Linear and Non-Linear Separation of Patterns by Linear Programming," Operations Research 13, 1965, pp. 444-452.

10. Mangasarian, O. L., "Multi-Surface Method of Pattern Separation," IEEE, Transactions on Information Theory, November 1968.

11. IBM Publication H20-0476-1, Mathematical Programming System 360, Linear and Separable Programming - Users Manual.

12. Simmonard, M., "Linear Programming," Prentice Hall, 1966.

13. Smith, F. W., "Pattern Classifier Design by Linear Programming," IEEE Transactions on Computers, Vol. C-17, No. 4, April 1968, pp. 367-372.